Exploring the Complex Folding Kinetics of RNA Hairpins: II. Effect of Sequence, Length, and Misfolded States

Wenbing Zhang and Shi-Jie Chen
Department of Physics and Astronomy and Department of Biochemistry, University of Missouri, Columbia, Missouri

ABSTRACT The complexity of RNA hairpin folding arises from the interplay between the loop formation, the disruption of the slow-breaking misfolded states, and the formation of the slow-forming native base stacks. We investigate the general physical mechanism for the dependence of the RNA hairpin folding kinetics on the sequence and the length of the hairpin loop and the helix stem. For example, 1), the folding would slow down when a stable GC basepair moves to the middle of the stem; 2), hairpin with GC basepair near the loop would fold/unfold faster than the one with GC near the tail of the stem; 3), within a certain range of the stem length, a longer stem can cause faster folding; and 4), certain misfolded states can assist folding through the formation of scaffold structures to lower the entropic barrier for the folding. All our findings are directly applicable and quantitatively testable in experiments. In addition, our results can be useful for molecular design to achieve desirable fast/slow-folding hairpins, hairpins with/without specific misfolded intermediates, and hairpins that fold along designed pathways.

INTRODUCTION

Recent experimental and theoretical studies on RNA hairpin folding kinetics are beginning to shed light on full complex folding energy landscapes and folding kinetics for RNA (and DNA) hairpins (1–18). RNA hairpin folding kinetics is found to give a wide range in magnitude and sign of the folding activation barriers for different sequences and for different temperatures (12–18). Furthermore, from the general theory developed in the previous article, we find that RNA hairpins, even though simple in structure, can be very complex in the folding kinetics. For example, depending on the nucleotide sequence, the folding can be rate-limited by the formation of the loop; or by the slow formation of the base stacks; or by the slow disruption of a misfolded non-native base stack. Moreover, the hairpin structure can form cooperatively through a two-state transition, or noncooperatively through multiple intermediate states. And a decrease in temperature can accelerate or decelerate the folding process.

Different sequences of the hairpins can have a wide range of very different folding kinetics behaviors. Most of the previous studies are focused on isolated sequences and the effect of loop closure on the folding kinetics. In this study, we go beyond the isolated sequences by exploring systematically the sequence and structural dependence of the folding kinetics by investigating how the loop length, loop sequence, stem length, and stem sequence affect the hairpin folding kinetics. In addition, we investigate the effect of the kinetic intermediates, especially the misfolded intermediates, on the folding kinetics. We found that certain misfolded intermediates may assist the folding process by lowering the entropic barrier of folding.

Since this study is based on the general RNA hairpin folding theory developed in the previous article, we first briefly summarize major conclusions from the general theory.

We describe the chain conformations according to base stacks. Different conformations are kinetically connected through a kinetic move set defined as the formation and disruption of a base stack or a stacked basepair. The rate of a kinetic move is given by $k_+ = k_0\, e^{-\Delta S/k_B}$ and $k_- = k_0\, e^{-\Delta H/k_B T}$ for the formation and breaking of a base stack (or a basepair), respectively. Here ΔS and ΔH are the corresponding entropy and enthalpy changes. As a result, the rate-limiting steps of folding correspond to the formation of the native base stacks with the largest entropy decrease ΔS and the disruption of the non-native base stacks with the largest enthalpy cost ΔH.

RNA hairpin folding can involve the following four types of rate-limiting steps:

1. *Loop nucleation*, i.e., the formation of the first base stack of the chain. The process involves the entropy loss from loop closure as well as the formation of the base stack that closes the loop. The rate constant of loop closure is $k_{loop} = k_0\, e^{-(\Delta S_{loop} + \Delta S_{stack})/k_B}$, where ΔS_{stack} and ΔS_{loop} are the corresponding entropy losses.
2. *Formation of the rate-limiting stack*. The formation of certain base stack s^* may involve a significantly large entropy loss $\Delta S_{stack}{}^*$ and thus has a slow rate:

$$k_f^* = k_0\, e^{-\Delta S^*_{stack}/k_B}. \quad (1)$$

3. *Direct folding*. If the loop is closed by a rate-limiting (slow) stack s^*, the loop closure would be extremely slow with a rate constant of

$$k_{direct} = k_f^*\, e^{-\Delta S_{loop}/k_B} \ll k_f^*. \quad (2)$$

4. *Detrapping*. The disruption of non-native (*nn*) base stacks has a rate constant of $k_{detrap} \sim k_0\, e^{-\Delta H_{nn}/k_B T}$, where ΔH_{nn} is

Submitted March 14, 2005, and accepted for publication September 30, 2005.
Address reprint requests to Shi-Jie Chen, E-mail: chenshi@missouri.edu.

0006-3495/06/02/778/10 $2.00
doi: 10.1529/biophysj.105.062950

the enthalpy cost for the disruption of the non-native base stack. k_{detrap} is slow for large ΔH_{nn} or low temperature T.

If there is one rate-limiting base stack s^*, according to the possible rate-limiting steps, we can classify the conformational ensemble into five types of clusters (i.e., $C, N_{\text{n}}, N_{\text{nn}}, I_{\text{n}}$, and I_{nn}):

$$\begin{aligned} C &= \text{the fully unfolded conformation that contains no base stack;} \\ N_{\text{n}} + N_{\text{nn}} = N &= \text{conformations with the rate-limiting stack } s^* \text{ formed;} \\ I_{\text{n}} + I_{\text{nn}} = I &= \text{all other conformations (without } s^* \text{ formed).} \end{aligned} \tag{3}$$

Here the subscripts n and nn denote the conformations without and with the non-native stacks (in the respective clusters), respectively. If k_{detrap} is large, I_{n} and I_{nn} in cluster I can equilibrate quickly, resulting in a merged cluster $I = I_{\text{n}} + I_{\text{nn}}$, and similarly, N_{n} and N_{nn} in cluster N merge into a pre-equilibrated cluster $N = N_{\text{n}} + N_{\text{nn}}$.

The folding kinetics is a result of the intercluster transitions. In a cluster, there are two types of conformations: pathway conformations and nonpathway conformations. Conformations that directly participate in the intercluster transitions are called *pathway conformations*. All other conformations are *nonpathway conformations*. Therefore, the intercluster transitions (between cluster U and N) are realized by the kinetic moves between the pathway conformations U_{i} in cluster U and the pathway conformations N_{i} in cluster N, and the resultant rate constant is given by

$$k_{\text{U}\to\text{N}} = \sum_{\text{i}} [U_{\text{i}}] k_{\text{U}_{\text{i}}\to\text{N}_{\text{i}}};\ k_{\text{N}\to\text{U}} = \sum_{\text{i}} [N_{\text{i}}] k_{\text{N}_{\text{i}}\to\text{U}_{\text{i}}}, \tag{4}$$

where $[U_{\text{i}}]$ and $[N_{\text{i}}]$ are the equilibrium fractional populations (i.e., Boltzmann distribution) of U_{i} and N_{i} in the respective clusters. The kinetic partitioning factor (equal to the probability for taking a microscopic pathway, e.g., $U_{\text{i}} \to N_{\text{i}}$) is determined by

$$f_{\text{i}}^{(\text{path})} = \frac{[U_{\text{i}}] k_{\text{U}_{\text{i}}\to\text{N}_{\text{i}}}}{k_{\text{U}\to\text{N}}}. \tag{5}$$

The pathways with the largest $f_{\text{i}}^{(\text{path})}$ are the dominant pathways for $U \to N$. Higher stability (larger $[U_{\text{i}}]$ in Eq. 4) of the pathway conformations (versus nonpathway conformations) and higher stability of the fast-rate pathway conformations (larger $k_{\text{U}_{\text{i}}\to\text{N}_{\text{i}}}$ in Eq. 4) result in a faster kinetics.

Depending on the nucleotide sequence, the Arrhenius plot of the rate-temperature dependence can show non-Arrhenius behavior: there exists a rollover temperature T_{r} such that the folding activation barrier changes from positive for $T \le T_{\text{r}}$ to negative for $T > T_{\text{r}}$, and the folding kinetics changes from noncooperative (multi-state) to cooperative. Summarized in Table 1 are the four folding kinetic scenarios in different temperature regimes.

Loop-length dependence

In this section, we investigate the loop-length dependence of the folding kinetics. To be specific, we study a series of sequences UAUAUCGC$_{\text{n}}$CGAUAUA(n = 3–9). The sequences have different loop lengths but have the same helix stem in the native structure. Moreover, the sequences have the same rate-limiting base stack s^* = (U,C,G,A), which has the largest ΔS and ΔH (see Fig. 1). We find that as loop size is increased, the folding rate decreases, but the unfolding rate nearly does not change (data not shown). Moreover, we find that the folding rate k_{f} scales with the loop size n as $k_{\text{f}} \sim n^{-1.8}$ at T = 30°C (Fig. 2). These findings agree with the experimental measurements for hairpin-folding kinetics (18).

To understand the loop-length dependence, we consider the cooperative folding condition (scenario 2: $T > T_{\text{r}} \simeq 10$°C) and use the two-cluster model with the native cluster N and unfolded cluster $U = C + I$. We first consider the unfolding transition $N \to U$ for the breaking of the rate-limiting stack s^*. The rate $k_{\text{N}\to\text{U}}$ is given by the sum over all the pathway conformations $\sum_{\text{i}} [N_{\text{i}}] k_{\text{N}_{\text{i}}\to\text{U}_{\text{i}}}$. Because both $[N_{\text{i}}]$ (= the fractional population of N_{i}) and $k_{\text{N}_{\text{i}}\to\text{U}_{\text{i}}}$ are independent of the loop length n, the unfolding rate is independent of the loop size.

The folding transition $U \to N$ corresponds to the formation of s^*. The rate $k_{\text{U}\to\text{N}}$ is given by $\sum_{\text{i}} [U_{\text{i}}] k_{\text{U}_{\text{i}}\to\text{N}_{\text{i}}}$. Except the direct folding pathway $U_1 \to N_1$, which has an extremely small rate k_{direct} (see Eq. 2), the other 19 pathways have the rate $k_{\text{U}_{\text{i}}\to\text{N}_{\text{i}}} \sim k_{\text{f}}^* \sim k_0 e^{-35.5/k_{\text{B}}} = 1.29 \times 10^6$ (s^{-1}) for the formation of s^*. The fractional population $[U_{\text{i}}]$ ($i > 1$) depends on the loop size through $[U_{\text{i}}] \sim e^{-\Delta S_{\text{loop}}/k_{\text{B}}}$. So $k_{\text{f}} \sim e^{-\Delta S_{\text{loop}}/k_{\text{B}}}$, where ΔS_{loop} is the entropy of the native hairpin loop. From the experimental measurements (19) and the theoretical modeling (20), the loop entropy is $\Delta S_{\text{loop}} \sim k_{\text{B}} \ln n^{-1.8}$. So we have $k_{\text{f}} \sim n^{-1.8}$ (see Fig. 2). This scaling law for the folding rate, which is obtained from the kinetic cluster analysis, agrees nearly exactly with the experimental data (18).

TABLE 1 A summary for the different scenarios of the folding kinetics

Scenario	Rate-limiting step	Cooperativity	Temperature
1	Loop formation	Two-state; $C \to F\ (= I + N)$	$T > T_{\text{r}}$
2	Formation of the rate-limiting native base stack s^*	Two-state; $U\ (= C + I) \to N$	$T > T_{\text{r}}$
3	Formation of s^* and detrapping from the non-native states	Multi-state; C, I_{n}, I_{nn}, N_{n}, N_{nn}	$T \le T_{\text{r}}$
4	Rate-limiting steps not discrete	Glassy	$T < T_{\text{r}}$

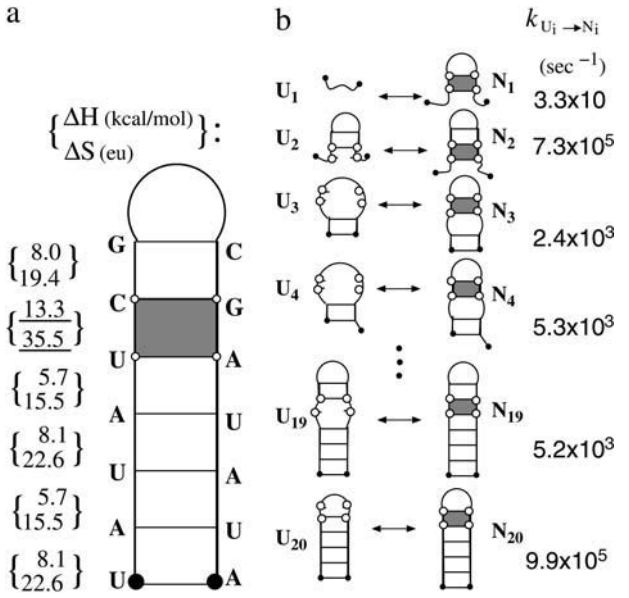

FIGURE 1 (*a*) The native structure and enthalpic and entropic parameters of the native state for sequence UAUAUCGC$_n$CGAUAUA. The shaded stack is the rate-limiting stack *s**. (*b*) The pathway conformations U_i and N_i ($i = 1, 2, \ldots, 20$) in the respective clusters *U* and *N*, the corresponding intercluster pathways $U_i \leftrightarrow N_i$, and the rate constants.

In the present model, the unfolding is rate-limited by the disruption of the rate-limiting stack *s**. Since the enthalpy cost ΔH^* for breaking *s** is assumed to be *n*-independent (under 1 M NaCl condition), the unfolding rate $k_u \sim e^{-\Delta H^*/k_B T}$ would be nearly independent of the loop size *n*. However, for small loops under lower ionic concentrations, the loop can be stabilized by excess loop-stem interaction (21–23). Considering the *n*-dependence of such excess stabilization ΔH_{excess}, the unfolding rate $k_u \propto e^{-\Delta H_{excess}/k_B T}$ can be *n*-dependent. Specifically, the loop would unfold faster for larger *n*. In fact, the *n*-dependence of the unfolding rate has been estimated from experiments as $k_u \sim n^{2.3}$ for DNA hairpins under 0.1 M NaCl (18). However, k_u for RNA hairpin folding (in 1 M NaCl) may scale differently.

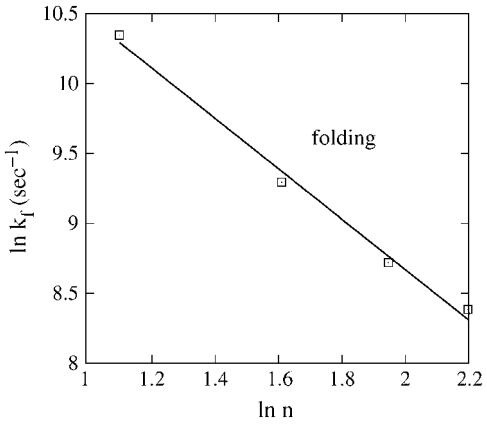

FIGURE 2 Loop length-dependence of the relaxation rate for sequence (UAUAUCGC$_n$CGAUAUA) at $T = 30$°C. Symbols ($n = 3, 5, 7, 9$): the folding rate solved from the exact master equation for the original (unclustered) conformational ensemble. (*Line*) The scaling law $k_f \sim n^{-1.8}$.

Stem-length dependence

In this section, we investigate the stem-length dependence of folding rate. By adding AU or UA basepairs to the helical stem of the sequence with $n = 5$ in the previous section, we generate a series of sequences with the same loop size but different stem length: $(AU)_m CGC_5 CG(AU)_m$ ($m = 2, 3, \ldots$). As shown in Fig. 3, we find the stem-length dependence of the folding and unfolding rate, as discussed below.

Cooperative folding regime ($T_r < T < T_m$; scenario 2 in Table 1)

Here $T_m \sim 50$°C is the melting temperature (computed from the statistical mechanical model (20)) and $T_r \sim 10$°C is the rollover temperature. The fast-folding pathway conformations in cluster *U* contain helical stems (see U_{19} and U_{20} in Fig. 1 *b*), and the longer helix stem enhances the stability of these fast-folding conformations. Therefore, a longer stem leads to faster folding. However, if the stem is too long, the nonpathway conformations (non-native states in I_{nn}) can be very stable and can dominate the population. This would effectively destabilize the pathway conformation and cause a slow folding.

Noncooperative folding ($T < T_r$; scenarios 3 and 4 in Table 1)

In this case, detrapping is rate-limiting. As the chain is elongated, the number of non-native conformations quickly

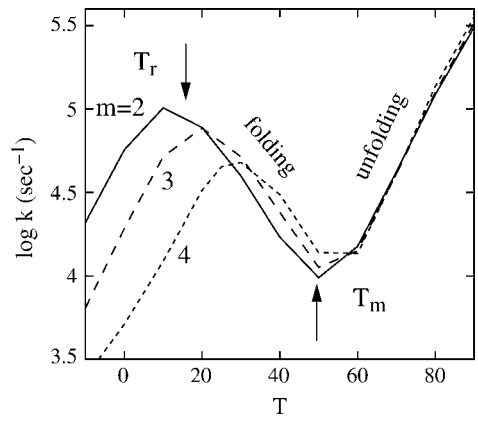

FIGURE 3 Temperature (T in °C) and stem-length dependence of the relaxation rate k_r for $(AU)_m CGC_5 CG(AU)_m$ with $m = 2$ (*solid line*), $m = 3$ (*dashed line*), and $m = 4$ (*dotted line*). The rollover temperature (T_r) and melting temperature (T_m) are different for the sequences. At $T < T_r$, the k_r-value decreases as T is decreasing; at $T_r < T < T_m$, the k_r-value increases as T is decreasing; and at $T > T_m$, the k_r-value increases as T is increasing.

increases. This greatly enhances the probability for the chain to fold to the misfolded states, causing a slower folding.

Cooperative unfolding ($T > T_m$; scenario 2 in Table 1)

At the unfolding temperature $T > T_m$, the dominant kinetic process is the unfolding. The rate is determined by the (unfolding) rate of the disruption of the rate-limiting stack ($N \rightarrow U$),

$$k_u = \sum_{i=1}^{20} [N_i] k_{N_i \rightarrow U_i}.$$

Here, $k_{N_i \rightarrow U_i} \sim k_0 e^{-\Delta H^*/k_B T}$ and ΔH^* is the enthalpy of the rate-limiting stack (U,C,G,A). Since the stem length only weakly affects the fractional population $[N_i]$ of N, the unfolding rate k_u is independent of the stem length.

Loop-sequence dependence

The loop sequence can affect the folding kinetics through two effects: (1), the sequence-dependent, single-stranded stacking in the loop region; and (2), the possible formation of non-native basepairs between the loop and the stem. Here we explore the loop-sequence dependence due to the formation of the non-native basepairs. We make a loop mutation $C^{12} \rightarrow G$ for sequence $(AU)_2CGAUAC_5UAUCG(AU)_2$ (see Fig. 4). The mutation does not alter the native structure (shown in Fig. 4 *a*) and the unfolding rate, but it notably changes the folding rate and its temperature-dependence (see Fig. 5 *a*): (1), the wild-type sequence folds much faster than the mutant sequence; and (2), they show opposite temperature-dependence: as the temperature is increased, the wild-type folds more slowly and the mutant sequence folds more quickly.

The wild-type sequence has two rate-limiting stacks: $s_1^* = (4, 5, 19, 20) = (U, C, G, A)$; and $s_2^* = (6, 7, 17, 18) = (G, A, U, C)$ (see Fig. 4). The formation of s_1^* and s_2^* have rate constants of $k_{f1}^* = 1.3 \times 10^6\, s^{-1}$ and $k_{f2}^* = 1.3 \times 10^6\, s^{-1}$, respectively. According to the two rate-limiting stacks, we classify the conformational ensemble into four clusters:

$$\begin{aligned} U &= \text{states with neither } s_1^* \text{ nor } s_2^*; \\ I_1 &= \text{states with } s_1^* \text{ and without } s_2^*; \\ I_2 &= \text{states with } s_2^* \text{ and without } s_1^*; \\ N &= \text{states with both } s_1^* \text{ and } s_2^*. \end{aligned} \quad (6)$$

To be specific, we study the kinetics at a representative temperature $T = 40°C$. We construct the 4×4 rate matrix for the four-cluster system (see Fig. 6). The eigenvalues of the four-cluster system are (0, 4.03×10^3, 5.96×10^5, 8.46×10^5) s^{-1}. The large gap between the lowest nonzero rate and the next nonzero rate clearly indicates that the folding process is single-exponential and the overall folding rate is $4.03 \times 10^3\ s^{-1}$. How can the two rate-limiting steps result in a single-exponential kinetics?

1. The formation of the first rate-limiting stack (s_1^* through $U \rightarrow I_1$ or s_2^* through $U \rightarrow I_2$) is extremely slow and is the bottleneck for the overall folding process. The rate is slow because in cluster U, the most populated state (= the fully unfold state) is slow-folding (through direct folding), with the extremely small rate k_{direct} (see Eq. 2), while the fast-folding conformations (i.e., stacked conformations) occupy <1% of total population in U.
2. With the first rate-limiting stack formed, the pathway conformations in cluster I_1 and I_2 would further fold through the formation of the second rate-limiting stack with rate k_{f1}^* for s_1^* or k_{f2}^* for s_2^*. Both k_{f1}^* and k_{f2}^* are much faster than the rate for the formation of the first stack. Therefore, the overall folding is rate-limited by the formation of the first stacks and the resultant folding kinetics is single-exponential with a rate of $k_f = k_{U \rightarrow I_1} + k_{U \rightarrow I_2}$. Equation 4 gives $k_{U \rightarrow I_1} = 2.82 \times 10^3 s^{-1}$ and $k_{U \rightarrow I_2} = 1.65 \times 10^3 s^{-1}$, so $k_f = 4.47 \times 10^3\ s^{-1}$, which is very close to the result from the rigorous eigenvalue $4.03 \times 10^3\ s^{-1}$. In Fig. 4 *b*, we show the dominant pathways predicted from the kinetic partitioning factor f_i^{path} (see Eq. 5) in the kinetic cluster analysis. As temperature is increased, the slow-folding (fully unfolded) state in U is stabilized, causing a decrease in the folding rate.

What causes the drastically different folding kinetics for the loop mutation? The mutation causes the stabilization of the (nonpathway) misfolded conformations in cluster U. For example, at $T = 30°C$, the loop mutation causes the nonpathway conformation population in U to increase from 44.3% (for the wild-type) to 91.7%. Such a dramatic change is due to the formation of stable non-native structures (see Fig. 5 *b*) formed by the basepairing between a G in the loop and a C in the stem. Stabilizing the nonpathway conformations effectively destabilizes the pathway conformations and causes a decrease in the folding rate. Higher T would destabilize this misfolded state (population drops from 91.7% to 70% as T increases from 30°C to 40°C) and effectively

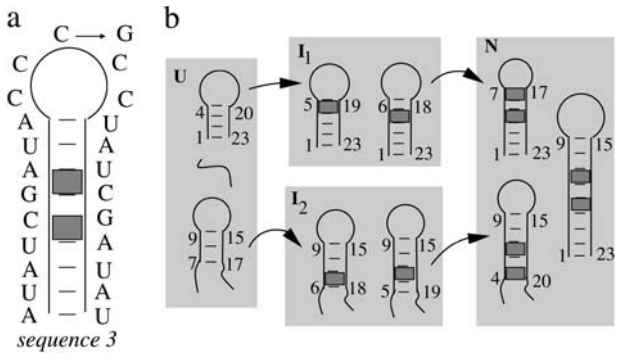

FIGURE 4 (*a*) The wild-type and mutant sequence and structure. (*b*) The four-cluster system and the most probable folding pathways for U (unfolded) → N (folded) for the wild-type sequence.

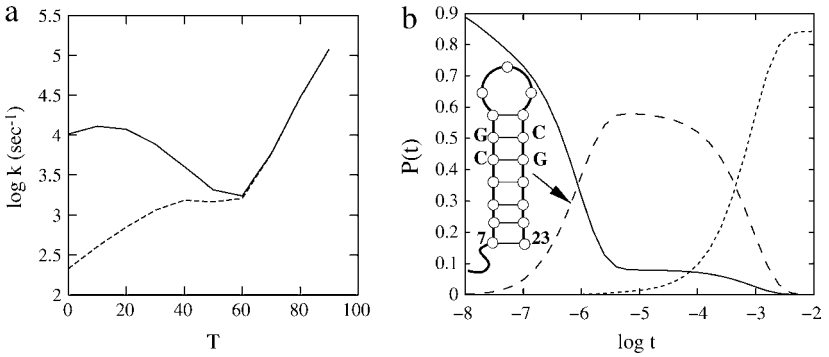

FIGURE 5 (*a*) The temperature (T in °C) dependence of the relaxation rate for the wild-type sequence (*solid line*) and the mutant sequence (*dashed line*). (*b*) The populational kinetics of the denatured state (*solid line*), the intermediate state (*long dashed line*), and the native state (*short dashed line*) for the mutant sequence at $T =$ 30°C. The inset is the structure of the misfolded intermediate state. Time t is in units of seconds.

stabilizes the pathway conformations in cluster U and causes a faster folding.

Is this misfolded state a kinetic trap that prevents the pre-equilibration process? No. In fact, it is the result of the pre-equilibration of cluster U. The emergence of the transient intermediate is due to its low free energy relative to all the other states in cluster U. Because its free energy is high relative to the states in N, the intermediate exists only transiently and would disappear when the chain folds into cluster N and the system relaxes to the final equilibrium state.

Stem-sequence dependence

In this section, we study three sequences that have the same loop size and the same stem length, but different stem sequences: sequences *1*, *2*, and *3*, which are shown in Fig. 7, *b* and *c*, and Fig. 4 *a* (wild-type), respectively. The three stem sequences differ by the different positions of two consecutive GC basepairs that form a stable (G, C, G, C) base stack as a clamp in the helix. Specifically, sequences *1*, *2*, and *3* have the GC clamp near the hairpin loop, at the tail of the stem, and in the middle of the stem, respectively. Sequences *1* and *2* contain one rate-limiting stack, and sequence *3* contains two rate-limiting stacks; see Fig. 7, *b* and *c*, and Fig. 4 *a*, respectively. Plotted in Fig. 7 *a* are the temperature-dependence of the rates. From the figure, we make the following two observations:

1. Sequence *1* (with the GC clamp close to the loop) folds faster than sequence *2* (with the GC clamp close to the stem tail). They both have only one rate-limiting stack, so their conformations can both be classified into two clusters U and N (scenario *2* in Table 1), corresponding to conformations with and without the rate-limiting stack formed, respectively. To be specific, we use $T = 30$°C for illustration. At $T = 30$°C, the most populated pathway conformation in cluster U, except the fully unfolded state, which is extremely slow-folding, is shown in Fig. 7, *b* and *c*, for sequences *1* and *2*, respectively. They are the dominant folding pathways with $f^{(\mathrm{path})} = 91.8\%$ and 84.9%, respectively. The folding rates along these dominant pathways are $k_{\mathrm{seq2}} = k_0\, e^{-\Delta S^*/k_B}$ for sequence *2* and $k_{\mathrm{seq1}} = k_0\, e^{-(\Delta\Delta S_{\mathrm{loop}} + \Delta S^*)/k_B} = k_{\mathrm{seq2}}\, e^{-\Delta\Delta S_{\mathrm{loop}}/k_B} > k_{\mathrm{seq2}}$ for sequence *1*, where ΔS^* is the entropy change for the formation of the rate-limiting stack and $\Delta\Delta S_{\mathrm{loop}}$ is the entropy change due to the change of the loop size from length 7 to 5 in Fig. 7 *b*. $\Delta\Delta S_{\mathrm{loop}}$ is negative. So $k_{\mathrm{seq1}} > k_{\mathrm{seq2}}$, i.e., sequence *1* folds faster than sequence *2*.
2. As the GC clamp moves to the middle, the folding slows down. Sequence *3* has two rate-limiting stacks. As we discussed in the previous section, the folding is limited by the formation of the first rate-limiting stack. The corresponding dominant pathways for sequences *1* and *2* and for sequence *3* are shown in Fig. 7, *b* and *c*, and Fig. 4 *b*, respectively. The dominant pathway conformations in cluster U in Fig. 7, *b* and *c* (for sequences *1* and *2*), contain continuous stable stacks. However, such highly stacked pathway conformations are not possible for sequence *3* because the otherwise continuous base stacks would be disrupted by the (to-be-formed) rate-limiting stacks in the middle of the stem. As a result, the dominant pathway conformations for sequences *1* and *2* are more stable and the resultant folding rates are larger.

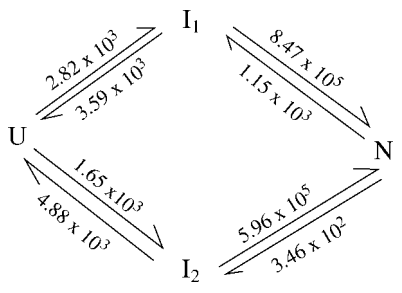

FIGURE 6 The intercluster transition rates (s^{-1}) at $T = 40$°C for the four kinetic clusters shown in Fig. 4 *b* for the wild-type sequence shown in Fig. 4 *a*.

Non-native structure-assisted RNA hairpin folding

For RNA hairpins, the formation of certain misfolded states can assist instead of delay the hairpin-folding process. We use hairpin-forming sequence AUAUCGAGAUCACCCU-CUCGAUAU to illustrate this. There are 1021 states for the sequence. The thermal denaturation for this sequence occurs at melting temperature $T_m = 68$°C (computed from the

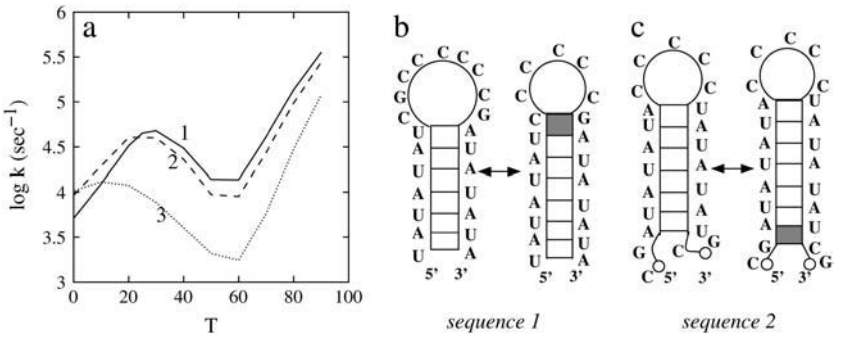

FIGURE 7 (*a*) The temperature (T in °C) dependence of the relaxation rate for the three sequences with the GC pair at different positions in the stem. The folding processes are rate-limited by the formation of the rate-limiting stack (*solid stack*). Panels *b* and *c* show the dominant folding pathways for sequences *1* and *2*, respectively.

statistical thermodynamics model (20)). We focus on the kinetics at $T = 40°C < T_m$.

To understand the microscopic folding pathways, we use the kinetic-cluster analysis. For this sequence, there are three slow-forming native base stacks (with large ΔS),

$$s_1^* = (4, 5, 19, 20) = (U, C, G, A);$$
$$s_2^* = (6, 7, 17, 18) = (G, A, U, C);$$
$$s_3^* = (8, 9, 15, 16) = (G, A, U, C), \quad (7)$$

and two slow-disruption non-native rate-limiting stacks (with large ΔH),

$$s_{1'}^* = (6, 7, 15, 16) = (G, A, U, C);$$
$$s_{2'}^* = (8, 9, 17, 18) = (G, A, U, C). \quad (8)$$

According to the rate-limiting stacks, we classify the conformational ensemble into 12 clusters:

U = the states without any of the rate-limiting stacks formed,

N = the states with all the rate-limiting native stacks formed,

and

$$I_1, I_2, I_3, I_{1'}, I_{2'}, I_{12}, I_{13}, I_{23}, I_{11'}, I_{12'}.$$

Here I_i = the states with s_i^* formed and I_{ij} = the states with both s_i^* and s_j^* formed. The eigenvalues of the 12-state kinetic cluster system are (0, 1.13, 2.38, 3.93, 5.93, 10.4, ...) $\times 10^4$ s^{-1}. The eigenvalue spectrum of the 12-state system agrees well with that of the original 1021-state system: (0, 1.09, 2.31, 3.80, 5.72, 10.2, ...) $\times 10^4$ s^{-1}. This validates our kinetic cluster analysis based on the 12-cluster system.

As we discussed for the folding with two (multiple) rate-limiting stacks, the formation of the first rate-limiting native stack is the bottleneck for the overall folding. From the kinetic connectivity diagram in Fig. 8 *a*, there exist two types of pathways for the formation of the first rate-limiting native stack (s_1, s_2, or s_3):

$$\text{On-pathway: } U \rightarrow I_i (i = 1, 2, 3) \text{ for the formation of } s_i^*;$$

and

$$\text{Off-pathway: } U \rightarrow I_{1'} \rightarrow I_{1i'} \rightarrow I_1 (i' = 1', 2').$$

So the total folding rate can be calculated as a sum of these (parallel) pathways:

$$k_f = \sum_{i=1,2,3} k_{U \rightarrow I_i} + \sum_{i'=1',2'} k_{U \rightarrow I_{i'} \rightarrow I_{1i'} \rightarrow I_1}. \quad (9)$$

In the above equation, $k_{U \rightarrow I_i}$ can be directly computed from Eq. 4. For $k_{U \rightarrow I_{i'} \rightarrow I_{1i'} \rightarrow I_1}$, considering the rebound from the two intermediate states I'_i and $I_{1i'}$, we have (24)

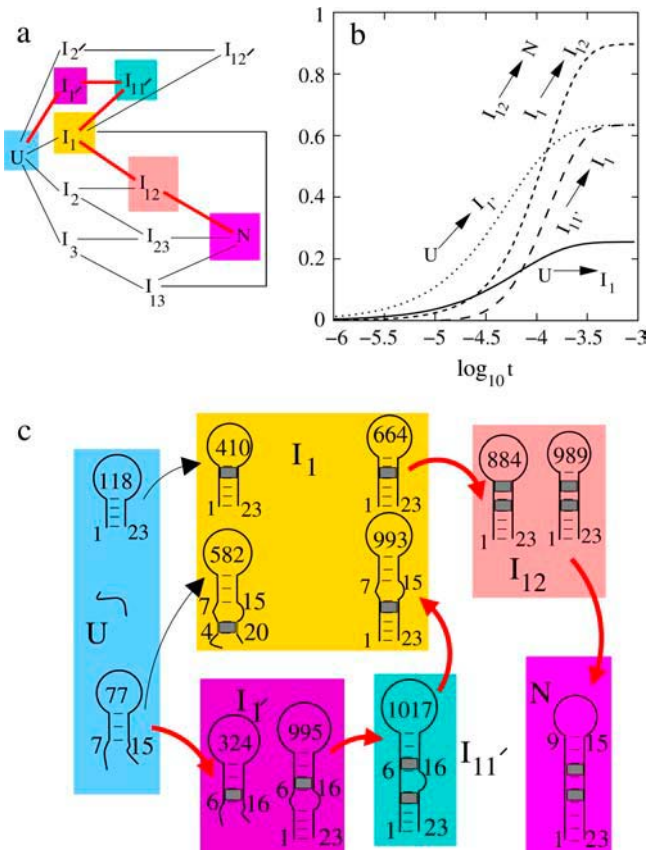

FIGURE 8 (*a*) The kinetic connectivity of the 12-cluster system (the *red lines* show the main folding pathway). (*b*) The net fluxes for the intercluster transitions. The net flux curve for $I_{12} \rightarrow N$ nearly coincides with the curve for $I_1 \rightarrow I_{12}$. This means that in the folding process, nearly all the chain conformations entering cluster I_{12} from I_1 would fold into the native cluster N. (*c*) The main pathways (in *red*) for the folding at $T = 40$°C.

$$k_{U\to I_{i'}\to I_{1i'}\to I_1} = (k_{U\to I_{i'}})(r_1 r_2)\sum_{n=0}^{\infty}[r_1(1-r_2)]^n, \qquad (10)$$

where

$$r_1 = \frac{k_{I_{i'}\to I_{1i'}}}{(k_{I_{i'}\to I_{1i'}} + k_{I_{i'}\to U})}$$

and

$$r_2 = \frac{k_{I_{1i'}\to I_1}}{(k_{I_{1i'}\to I_1} + k_{I_{1i'}\to I_{i'}})}$$

account for the rebound effect (see Fig. 9). Combining the above results, we have $k_f = 1.15 \times 10^4\ s^{-1}$. This k_f result, which is based purely on the intercluster pathway analysis, agrees very well with the first non-zero rate ($1.11 \times 10^4\ s^{-1}$) solved from the exact master equation for the original complete conformations ensemble.

Which pathway dominates the folding process, on-pathway or off-pathway? Because

$$\frac{k_{U\to I_{1'}}}{\sum_{i=1,2} k_{U\to I_i}\sum_{i=1,2,3} k_{U\to I_{i'}}} = 68\%$$

and

$$\frac{k_{U\to I_1}}{\sum_{i=1,2} k_{U\to I_i}\sum_{i=1,2,3} k_{U\to I_{i'}}} = 22.8\%,$$

only ~22.8% population in cluster U folds through the on-pathway route $U \to I_1$ and 68% folds through the off-pathway route $U \to I_{1'}$. Therefore, the folding is dominated by the off-pathway process.

To further characterize the populational statistics, we plot in Fig. 8 *b* the net populational fluxes along pathways $U \to I_{1'}$, $I_{11'} \to I_1$, $I_1 \to I_{12}$, and $I_{12} \to N$. The populational flux $P_{I\to J}$ is the (accumulated) probability for the molecule to fold through $I \to J$ during time period $0 \to t$. The populational flux from cluster I to cluster J is defined as (24):

$$P_{I\to J}(t) = \int_0^t (P_I(t')\times k_{I\to J} - P_J(t')\times k_{J\to I})dt',$$

where $P_i(t)$ is the population of the states in cluster i. The results in Fig. 8 *b* show that $P_{U\to I_{1'}} \gg P_{U\to I_1}$, and that $P_{I_{11'}\to I_1}$ and $P_{I_1\to I_{12}}$ quickly rise in the folding process, which confirms that the dominant pathway is the off-pathway route through the formation and disruption of the non-native base stack $s_1^{*\prime}$ ($U \to I_{1'} \to I_{11'} \to I_1 \to I_{12} \to N$). How does the formation of the non-native stack $s_1^{*\prime}$ in $I_{1'}$ facilitate the folding process?

From the unfolded state U, the formation of the non-native base stack $s_1^{*\prime}$ is much faster than the direct formation of the native base stack s_1^*. In the unfolded cluster U, except the fully unfolded state, which has negligible direct folding rate, the most stable pathway conformation is state 77 (see Fig. 8 *c*), which occupies 1.32% of the total population of U.

The dominant pathway for the formation of the native s_1^* is through $77 \to 582$. This pathway involves the closure of an internal loop, and thus has a slow rate of due to the entropic loss ($\Delta S_{intloop}$) for the formation of the internal loop closed by basepairs (4,20) and (7,15) in state 582 (see Fig. 8 *c*): $k_{77\to582} = k_0\, e^{-(\Delta S_1^* + \Delta S_{intloop})/k_B} = 4.16 \times 10^2\ s^{-1}$. Here ΔS_1^* is the entropy parameter for the formation of stack s_1^*.

On the other hand, the dominant pathway for the formation of the non-native $s_{1'}^*$ is through $77 \to 324$. Since this pathway does not involve the closing of additional loops, it has a much faster rate $k_{77\to324} = k_0\, e^{-\Delta S_{1'}^*/k_B} = 6.92 \times 10^5\ s^{-1}$.

So most of the population in U would quickly fold along the off-pathway route $77 \to 324$ to form the non-native rate-limiting stack $s_{1'}^*$.

Once the non-native base stack $s_{1'}^*$ is formed in state 324 in cluster $I_{1'}$, the pathway conformations in $I_{1'}$ can be quickly stabilized through the elongation of the helix stem (e.g., $324 \to 995$ in Fig. 8 *c*). These stabilized (non-native) pathway conformations would cause fast transitions from $I_{1'}$. In addition, the stable non-native structures in $I_{1'}$ can serve as scaffolds to lower the entropic barrier for the further formation of the native rate-limiting stack s_1^*. This would accelerate the folding process. For example, transition $995 \to 1017$ is accompanied by an entropic change $\Delta\Delta S_{intloop} < 0$ for the decrease in the internal loop size. As a result, $k_{995\to1017} = k_0 e^{-(\Delta S_1^* + \Delta\Delta S_{intloop})/k_B} = 5.42 \times 10^6\ s^{-1}$ is much faster than both the direct on-pathway folding rate $k_{77\to324} = 4.16 \times 10^2\ s^{-1}$ and the off-pathway rate $k_{77\to324} = 6.92 \times 10^5\ s^{-1}$.

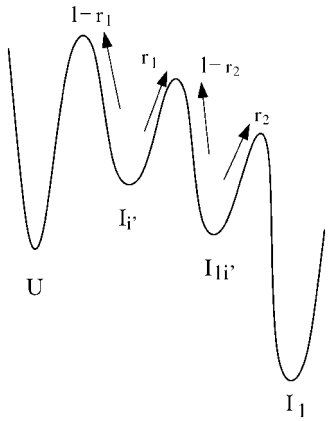

FIGURE 9 For the transition $U \to I_{i'} \to I_{1i'} \to I_1$, some of the population will rebound back from the intermediates $I_{i'}$ and $I_{1i'}$. The value r_1 is the probability from $U \to I_{i'}$, and $1 - r_1$ is the probability of rebound back from the intermediate from $I_{i'}$. The value r_2 is the rebound effect for the intermediate from I_{1i}'.

CONCLUSIONS

Although DNA and RNA hairpins are both stabilized by base-stacking interactions and both have loop formation as a slow step in the folding process, they can have very different folding kinetics. Unlike RNA hairpins, DNAs do not have large separations in the (ΔH_{stack}, ΔS_{stack}) parameters

for different base stacks. As a result, for most DNA sequences, hairpins fold through the formation of the stable loop (scenario 1) instead of the slow-folding native base stack (scenario 2).

Furthermore, the cluster model can explain the ion concentration-dependence of the folding and unfolding rates. Following Santalucia (25), we note that the enthalpy ΔH_{stack} for a base stack is nearly independent of $[Na^+]$, while the entropy is ΔS_{stack} for a base-stack decrease for higher $[Na^+]$ (25).

If the hairpin folding is rate-limited by the formation of a slow-forming base stack, the folding rate $k_f \sim e^{-\Delta S_{\text{stack}}/k_B}$ would increase as $[Na^+]$ is increased, while the unfolding rate $k_u \sim e^{-\Delta H_{\text{stack}}/k_B T}$ does not change with the ion concentration. These ion-dependences of k_f and k_u agree with the experimental results for RNA duplex association and dissociation kinetics (26).

If the hairpin folding is rate-limited by the loop formation (see Fig. 10 *a*), as the ion concentration is increased, the folding rate $k_f \sim e^{-(\Delta S_{\text{loop}}+\Delta S_{\text{stack}})/k_B}$ would increase due to the decrease in the entropy. The unfolding rate is given by $k_u \sim [c] e^{-\Delta H_{\text{stack}}/k_B T}$, where $[c]$ is the fractional population of state c in Fig. 10 *a*. Higher ion-concentration stabilizes structures with longer helix stems, e.g., state d (rather than state c) in Fig. 10 *a*, causing a smaller $[c]$ for state c, which has only one stack. As a result, k_u decreases as $[Na^+]$ is increased. Moreover, the temperature-dependence of k_u is dominated by the $e^{-\Delta H_{\text{stack}}/k_B T}$ factor, so the apparent activation barrier of the unfolding does not change with the ion concentration (ΔH_{stack} is assumed to be $[Na^+]$-independent). This is in agreement with the experimental finding (12).

The kinetic-cluster approach allows us to study the kinetic rates, rate-limiting steps, and the pathways for biologically significant RNA hairpins. In this study, we explore the sequence-dependent complex folding and unfolding kinetics for RNA hairpins. The overall hairpin folding process can be rate-limited by the formation of the loop, the formation of the rate-limiting native base stack, and the breaking of the stable non-native base stack. The competition between these different processes leads to the great wealth of different RNA hairpin-folding behavior. The detailed folding kinetics is sequence-specific. Our study reveals several intriguing features for RNA hairpin-folding kinetics (for $T > T_r$):

1. The unfolding rate is nearly independent of the loop-length n, and the folding rate decreases for larger loops and scales as $n^{-1.8}$.
2. For sequences with a rate-limiting native base stack, the high-temperature unfolding rate is relatively independent of the stem length. The folding rate increases for longer stem length due to the increased stability of the nativelike states. However, the folding rate would decrease if the stem is too long because of the formation of stable misfolded states.
3. The folding and unfolding kinetics can be dependent on the loop sequence. The basepairs between the loop region and the helical stem region can lead to stable misfolded kinetic intermediates and slow down the folding process. Especially, it is highly possible for the G (C) residues in the loop to pair with C (G) residues in the stem to form a stable non-native (G, C, G, C) stack.
4. The nucleotide sequence in the stem region is important for the folding/unfolding kinetics. For example, for a stem with GC pairs inserted in a series of AU pairs, the rate is larger for sequences with the GC basepairs close to the hairpin loop than for sequences with the GC pairs at the tail of the stem, and the rate decreases as the GC pairs move to the middle of the stem.
5. Folding can be assisted by the misfolded states because some stable misfolded states can be fast-folding by forming a scaffold structure to lower the entropic barrier for the formation of the native basepairs.

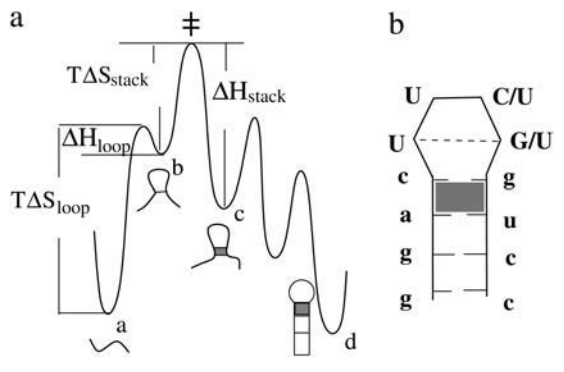

FIGURE 10 A schematic free energy landscape for hairpin folding and the native structure for ggacUUCGgucc (with tetraloop stabilization) or ggacUUUUgucc (without tetraloop stabilization).

These stem/loop length and sequence-dependence of the folding kinetics may be a paradigm for more complete and complex analysis of RNA folding kinetics. Moreover, the general length and sequence dependence can provide useful guidance for molecular design for folding rate, pathways, and cooperativity.

In this study, the effect of the specific loops such as the GNRA and UUCG tetraloops are not considered. These tetraloops can have excess stability due to the intraloop base stacking and hydrogen bonding (27–30). As shown below, it is possible to obtain a rough estimate for the kinetic effects by treating the tetraloop as a stable state (state b in Fig. 10 *a*) on the free energy landscape. To simplify the analysis, we use a rather crude energy landscape to represent the actual free energy landscape. Considering the rebound effect from the intermediate state b, we can estimate the forward folding rate k_f (24):

$$k_f \simeq k_{a\to b}\left(\frac{k_{b\to c}}{k_{b\to c}+k_{b\to a}}\right). \tag{11}$$

With the loop entropy ΔS_{loop} and enthalpy ΔH_{loop} for the tetraloop and the stacking entropy ΔS_{stack} for the (a, c, g, u)

stack (see the *shaded stack* in Fig. 10 *b*), our rate constant model gives $k_{a\to b} = k_0\, e^{-\Delta S_{loop}/k_B}$; $k_{b\to a} = k_0\, e^{-\Delta H_{loop}/k_B T}$; and $k_{b\to c} = k_0\, e^{-\Delta S_{stack}/k_B}$. The excess tetraloop stabilization parameter can be determined as $\Delta S_{excess} = \Delta S_{loop} - \Delta S_{loop}^{(0)}$ and $\Delta H_{excess} = \Delta H_{loop}$, where $\Delta S_{loop}^{(0)}$ is the entropy of the loop without the tetraloop stabilization.

To directly connect the theory to the experiment, we consider the YNMG RNA hairpins whose folding and unfolding rates have been measured by Proctor et al. (4). We specifically compare the folding rates for the following two sequences: ggacUUCGgucc (with tetraloop stabilization) and ggacUUUUgucc (without tetraloop stabilization). To extract the ΔS_{loop} and ΔH_{loop} for the experiment, we subtract the stem parameters from the experimentally measured hairpin parameters (4). Here the stem parameters are calculated from the Turner rule (19) with the salt corrections (with experimental condition of 10 mM Na^+) (25).

For the UUCG tetraloop, we found that $\Delta S_{excess} = 25$ eu and $\Delta H_{excess} = 12$ kcal/mol. Proctor et al. (4) measured that $k_f^{(exp)} = 6.1 \times 10^4\ s^{-1}$ at $T = 65°C$. Our theory (with Eq. 11) gives $k_f^{(model)} = 8.91 \times 10^4\ s^{-1}$, which is close to the experimental result. The unfolding rate can be estimated from the hairpin stability $\Delta G^{(exp)} = -0.79$ kcal/mol as $k_u \simeq k_f\, e^{\Delta G/k_B T}$, which gives $k_u^{(exp)} = 1.6 \times 10^4\ s^{-1}$ and $k_u^{(model)} = 2.3 \times 10^4\ s^{-1}$, respectively.

For the UUUU loop, there is no unusual tetraloop stabilization interaction. By assuming ΔH_{excess} and ΔS_{excess} to be zero in the above equations (i.e., $\Delta H_{loop} = 0$ and $\Delta S_{loop} = \Delta S_{loop}^{(0)}$), we found that $k_f^{(model)} = 2.13 \times 10^4 s^{-1}$ at $T = 65°C$, which is close to the experimental result $k_f^{(exp)} = 4.5 \times 10^4 s^{-1}$. The experimental and theoretical unfolding rates are $k_u^{(exp)} \simeq 12.8 \times 10^4 s^{-1}$ and $k_u^{(model)} \simeq 6.05 \times 10^4 s^{-1}$, respectively.

Consistent with the experimental finding, the theory predicts the acceleration in the folding process and the deceleration in the unfolding process due to the tetraloop stabilization. Physically, folding is accelerated because the excess intraloop stacking and basepairing can stabilize the transition state for the folding (see ‡ in Fig. 10 *a*) to lower the free energy barrier of folding. The unfolding is decelerated because the intraloop stacking and basepairing in the folded state can cause a higher (enthalpic) barrier for the disruption of the tetraloop.

We are grateful to Drs. Anjum Ansari and Herve Isambert for useful discussions.

This research was supported by the National Institutes of Health (NIH/NIGMS) through grant GM No. 063732 (to S.-J. C).

REFERENCES

1. Hall, K. B., and J. Williams. 2004. Dynamics of the IRE RNA hairpin loop probed by 2-aminopurine fluorescence and stochastic dynamics simulations. *RNA*. 10:34–47.
2. Jean, J. M., and K. B. Hall. 2001. 2-Aminopurine fluorescence quenching and lifetimes: role of base stacking. *Proc. Natl. Acad. Sci. USA*. 98:37–41.
3. Hall, K. B., and C. G. Tang. 1998. C-13 relaxation and dynamics of the purine bases in the iron responsive element RNA hairpin. *Biochemistry*. 37:9323–9332.
4. Proctor, D. J., H. Ma, E. Kierzek, R. Kierzek, M. Gruebele, and P. C. Bevilacqua. 2004. Folding thermodynamics and kinetics of YNMG RNA Hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry*. 43: 14004–14014.
5. Liphardt, J., B. Onoa, S. B. Smith, I. J. Tinoco, and C. Bustamante. 2001. Reversible unfolding of single RNA molecules by mechanical force. *Science*. 292:733–737.
6. Zhang, W. B., and S. J. Chen. 2002. RNA hairpin folding kinetics. *Proc. Natl. Acad. Sci. USA*. 99:1931–1936.
7. Zhang, W. B., and S. J. Chen. 2003. Master equation approach to finding the rate-limiting steps in biopolymer folding. *J. Chem. Phys.* 118:3413–3420.
8. Sorin, E. J., M. A. Engelhardt, D. Herschlag, and V. S. Pande. 2002. RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.* 317:493–506.
9. Sorin, E. J., Y. M. Rhee, B. J. Nakatani, and V. S. Pande. 2003. Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. *Biophys. J.* 85:790–803.
10. Sorin, E. J., B. J. Nakatani, Y. M. Rhee, G. Jayachandran, V. Vishal, and V. S. Pande. 2004. Does native state topology determine the RNA folding mechanism? *J. Mol. Biol.* 337:789–797.
11. Cocco, S., J. F. Marko, and R. Monasson. 2003. Slow nucleic acid unzipping kinetics from sequence-defined barriers. *Eur. Phys. J. E.* 10: 153–161.
12. Bonnet, G., O. Krichevsky, and A. Libchaber. 1998. Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc. Natl. Acad. Sci. USA*. 95:8602–8606.
13. Ansari, A., S. V. Kunznetsov, and Y. Shen. 2001. Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc. Natl. Acad. Sci. USA*. 98:7771–7776.
14. Kuznetsov, S. V., Y. Shen, A. S. Benight, and A. Ansari. 2001. A semiflexible polymer model applied to loop formation in DNA hairpins. *Biophys. J.* 81:2864–2875.
15. Wallace, M. I., L. Ying, S. Balasubramanian, and D. Klenerman. 2001. Non-Arrhenius kinetics for the loop closure of a DNA hairpin. *Proc. Natl. Acad. Sci. USA*. 98:5584–5589.
16. Wallace, M. I., L. Ying, S. Balasubramanian, and D. Klenerman. 2000. FRET fluctuation spectroscopy: exploring the conformational dynamics of DNA hairpin loop. *J. Phys. Chem. B*. 104:11551–11555.
17. Goddard, N. L., G. Bonnet, O. Krichevsky, and A. Libchaber. 2000. Sequence-dependent rigidity of single-stranded DNA. *Phys. Rev. Lett.* 85:2400–2403.
18. Shen, Y., S. V. Kunznetsov, and A. Ansari. 2001. Loop dependence of the dynamics of DNA hairpins. *J. Phys. Chem. B*. 105:12202–12211.
19. Serra, M. J., and D. H. Turner. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol.* 259:242–261.
20. Chen, S. J., and K. A. Dill. 2000. RNA folding energy landscapes. *Proc. Natl. Acad. Sci. USA*. 97:646–651.
21. Hilbers, C. W., C. A. Haasnoot, S. H. de Bruin, J. J. Joordens, G. A. van der Marel, and J. H. van Boom. 1985. Hairpin formation in synthetic oligonucleotide. *Biochimie*. 67:685–695.
22. Haasnoot, C. A., C. W. Hilbers, G. A. van der Marel, J. H. van Boom, U. C. Singh, N. Pattabiraman, and P. A. Kollman. 1986. On loop folding in nucleic acid hairpin-type structures. *J. Biomol. Struct. Dyn.* 3:843–857.
23. Groebe, D. R., and O. C. Uhlenbeck. 1988. Characterization of RNA hairpin loop stability. *Nucleic Acids Res.* 16:11725–11735.
24. Zhang, W. B., and S. J. Chen. 2003. Analyzing the biopolymer folding rates and pathways using kinetic cluster method. *J. Chem. Phys.* 119: 8716–8729.

25. SantaLucia, J. J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*. 95:1460–1465.

26. Porschke, D., O. C. Uhlenbeck, and F. H. Martin. 1973. Thermodynamics and kinetics of the helix-coil transitions of oligomers containing GC pairs. *Biopolymers*. 12:1313–1335.

27. Varani, G. 1995. Exceptionally stable nucleic acid hairpins. *Annu. Rev. Biophys. Biomol. Struct*. 24:379–404.

28. Pley, H. W., K. M. Flaherty, and D. B. McKay. 1994. Three-dimensional structure of a hammerhead ribozyme. *Nature*. 372:68–74.

29. Correll, C. C., and K. Swinger. 2003. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 resolution. *RNA*. 9:355–363.

30. Jager, J. A., D. H. Turner, and M. Zuker. 1989. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*. 86: 7706–7710.